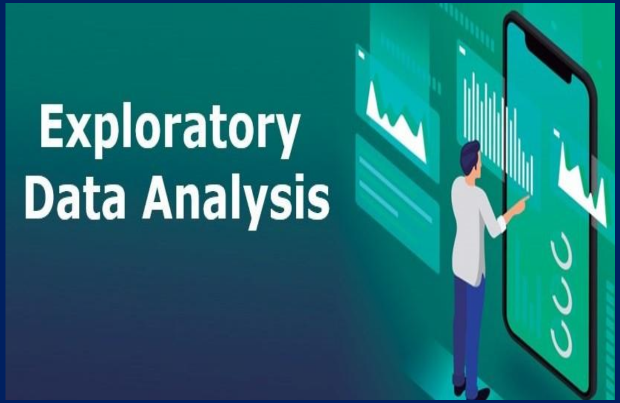
# BANK LOAN CASE STUDY



Using

Created by- Mainak Mukherjee
Email- subha.mainak@gmail.com

# PROJECT DESCRIPTION

This project focuses on evaluating the risk tolerance of banks when processing loan applications. The goal is to assist the bank in making informed loan approval decisions based on the applicant's profile. The bank faces two primary risks:

- If the applicant is likely to repay the loan but the loan is not approved, the bank misses out on potential business.
- If the applicant is unlikely to repay the loan (i.e., defaults), approving the loan could result in financial loss.

The provided data includes information about loan applications at the time they were submitted. There are two scenarios in the data:

1. Clients with payment difficulties: Applicants who were late by more than X days on at least one of the first Y installments.
2. All other cases: Applicants who made their payments on time.

A detailed analysis is required to extract insights from these scenarios. These insights will help the bank identify patterns and make decisions such as denying the loan, reducing the loan amount, or lending to high-risk applicants at a higher interest rate. This will ensure that applicants who are capable of repaying the loan are not unfairly rejected.

# APPROACH

- I utilized the COUNTA function to count the total number of rows in each column.
- Then, I calculated the percentage of null values in each column using the formula 1 - (Total Row Count for each column / Total Row Count).
- I removed all columns with more than 30% null values. For columns with less than 30% null values, I imputed the missing values using mean, median, and mode methods.
- Additionally, I identified outliers using the interquartile range method for relevant columns.
- After reviewing the description of each column, I retained only the relevant ones for analysis.
- Columns that represented days were converted into years by dividing the number of days by 365.

# TECH – STACK USED

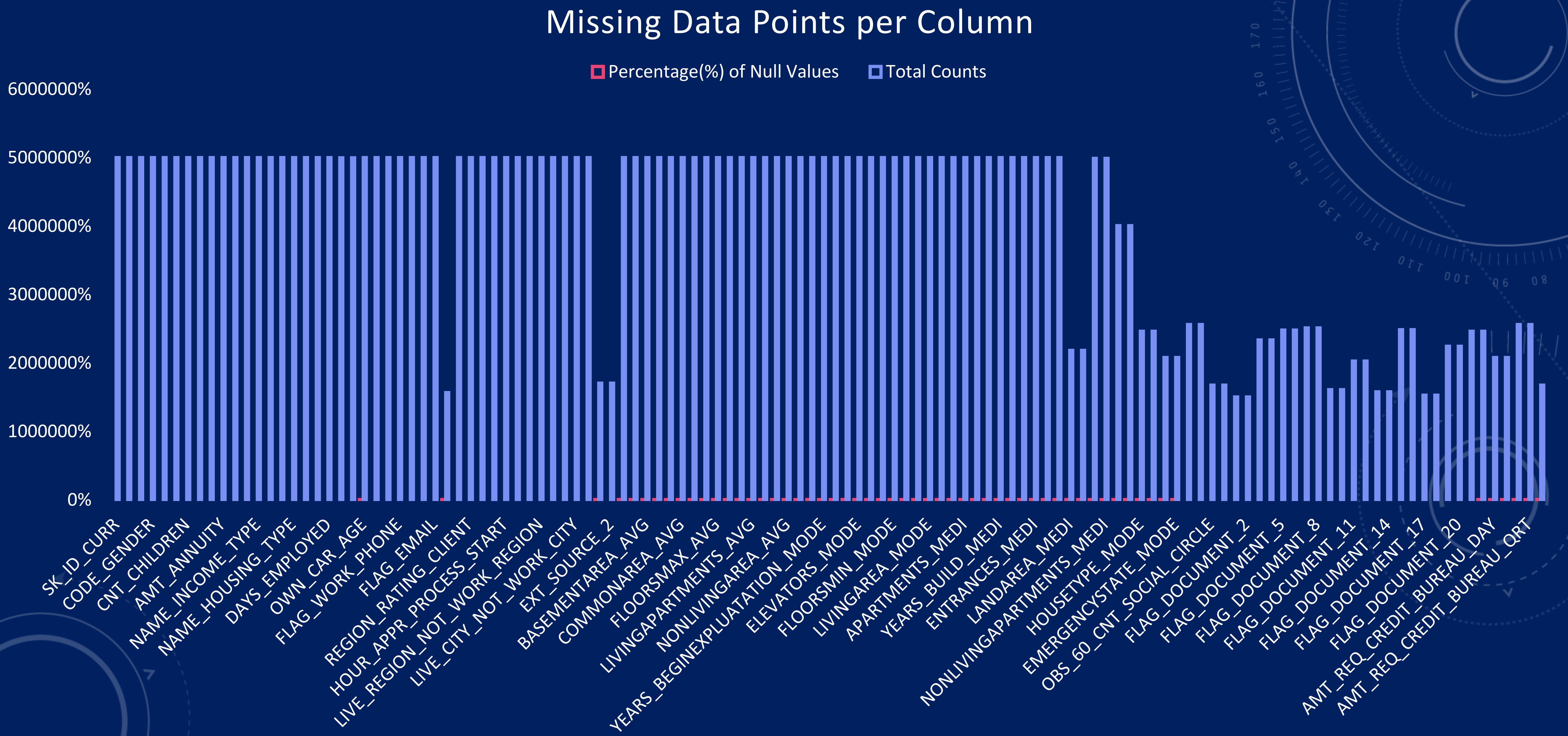
**For this project, the following tools were utilized:**

1. **MS Excel 2021:**
    - Used for data analysis.
    - Excel's built-in functions, formulas and tools were crucial for performing calculations.
    - Generated visualizations to illustrate data trends and patterns.
    - Summarized the results effectively.

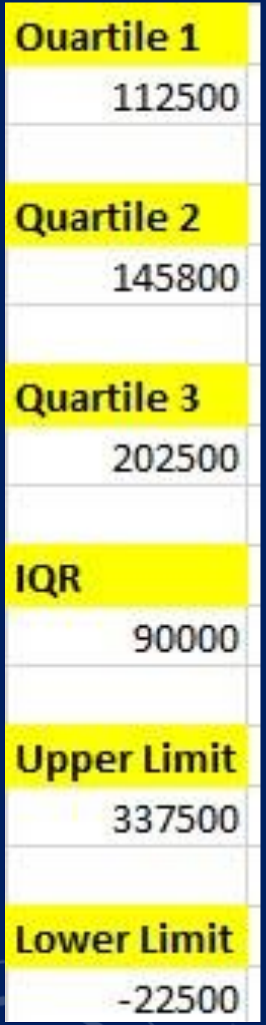
2. **MS PowerPoint 2021:**
    - Created a presentation to showcase project insights.
    - Ensured the information was presented in a clear and visually appealing format.

3. **Google Drive:**
    - Saved the final report for easy access.
    - Shared the report with others to facilitate collaboration and review.

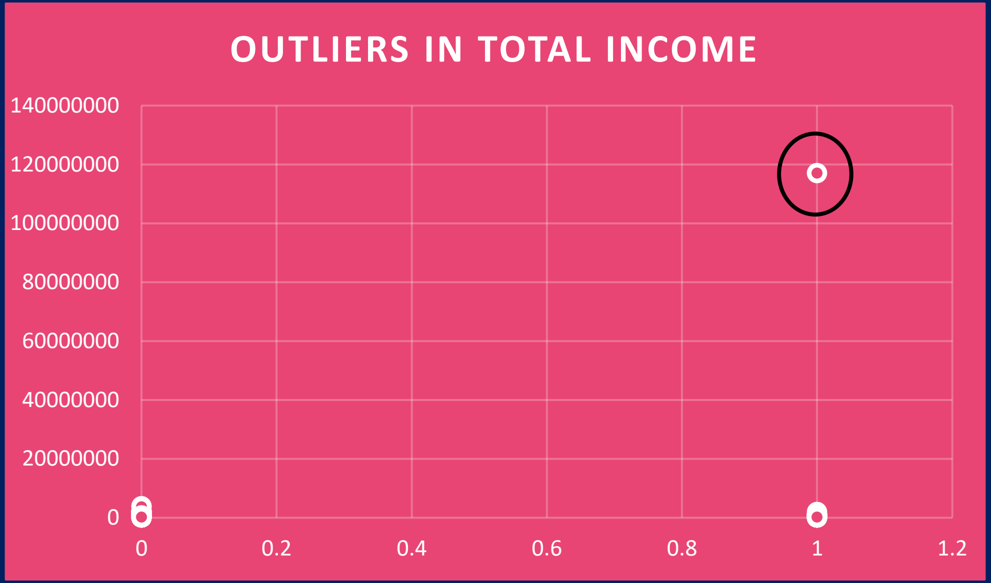
# Task 1:- Missing Data Identification

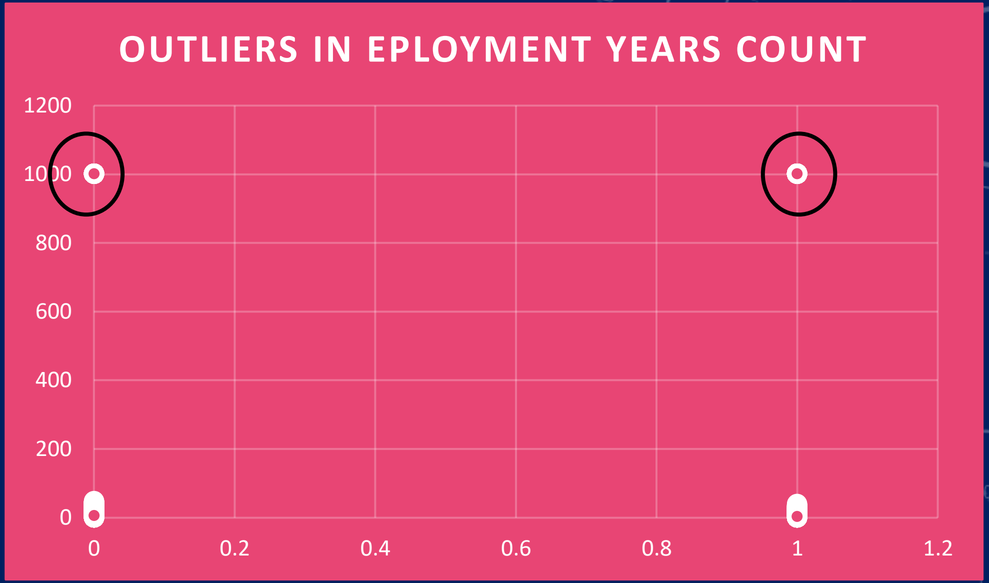


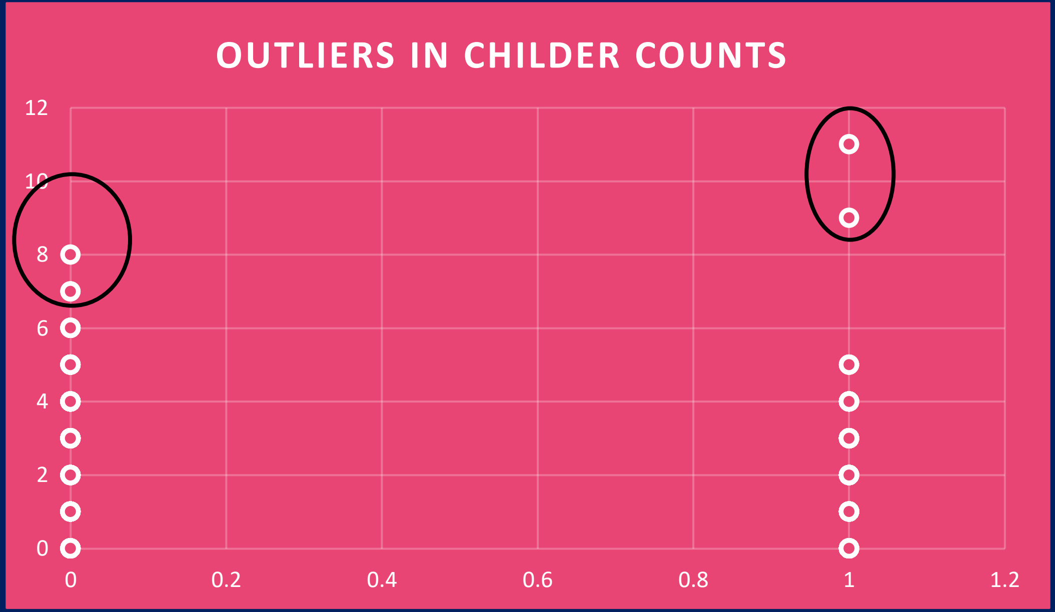
# Task 2:- Outliers Identification



| | |
|---|---|
| Quartile 1 | |
| | 112500 |
| Quartile 2 | |
| | 145800 |
| Quartile 3 | |
| | 202500 |
| IQR | |
| | 90000 |
| Upper Limit | |
| | 337500 |
| Lower Limit | |
| | -22500 |







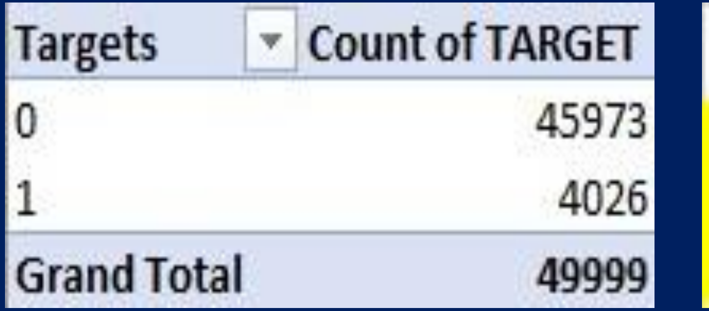
# Insights

**1. Outliers in Total Income:-**
For the target variable 1, some applicants have an income that is significantly higher than the usual range. Specifically, there are applicants with incomes around 11 crores, while the majority have incomes in the range of lakhs. For a detailed analysis, refer to the "outliers" columns visualization chart for AMT_TOTAL_INCOME.

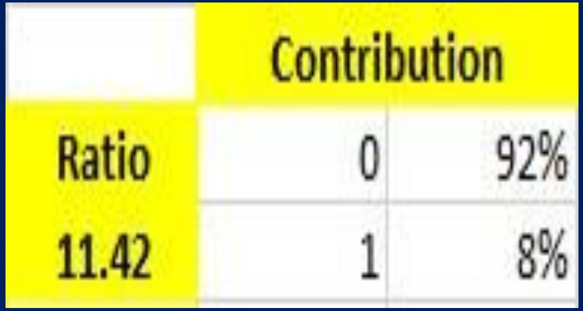
2. **Outliers in Employment Years Count:-**
In the "outliers" sheet for Days Employed, there are anomalies for both target columns 0 and 1. The XY plotter indicates some applicants have been employed for 1000 years as of the application date, which is clearly an error.

**3. Outliers in Children Counts:-**
In the "outliers" sheet for CNT_CHILDREN, there are anomalies in both target columns 0 and 1. The XY plotter for target 0 shows applicants with 8+ children, which is highly unusual in modern times. Similarly, the XY plotter for target 1 shows applicants with more than 10+ children.

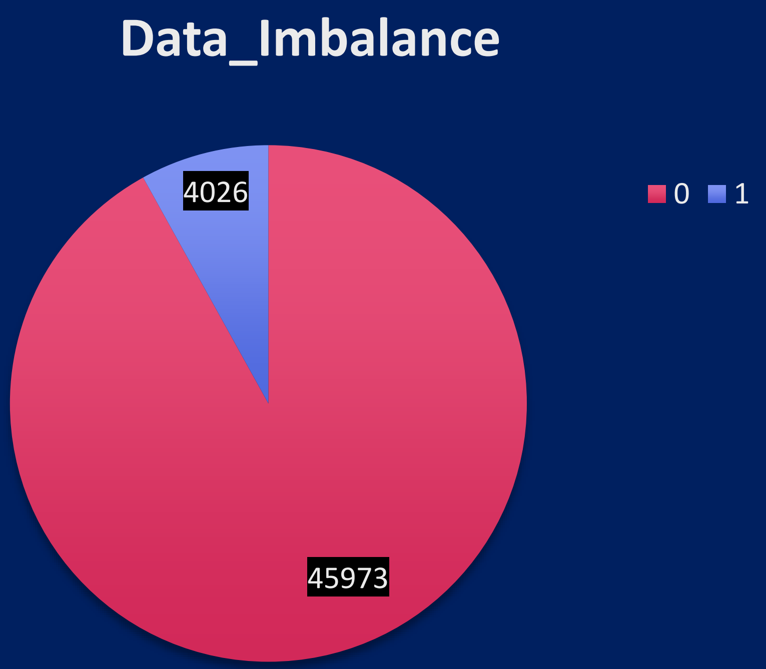
# Task 3:- Analyze Data Imbalance



| Targets | Count of TARGET |
|---|---|
| 0 | 45973 |
| 1 | 4026 |
| Grand Total | 49999 |



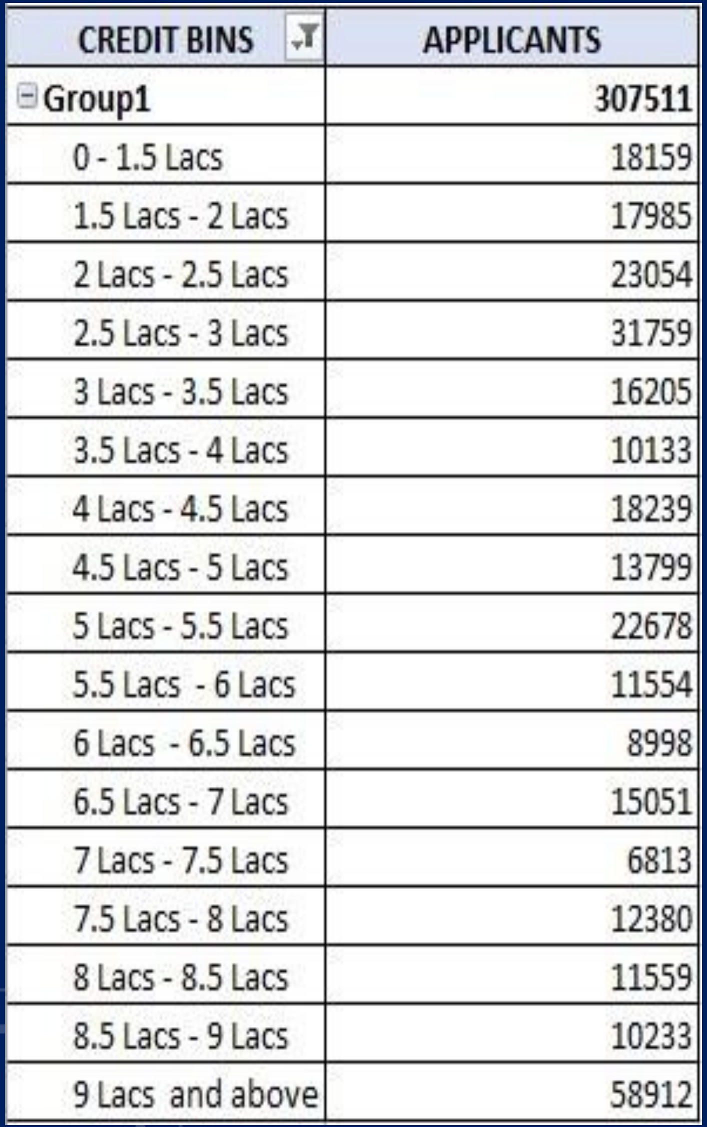
|  | Contribution |  |
|---|---|---|
| Ratio | 0 | 92% |
| 11.42 | 1 | 8% |



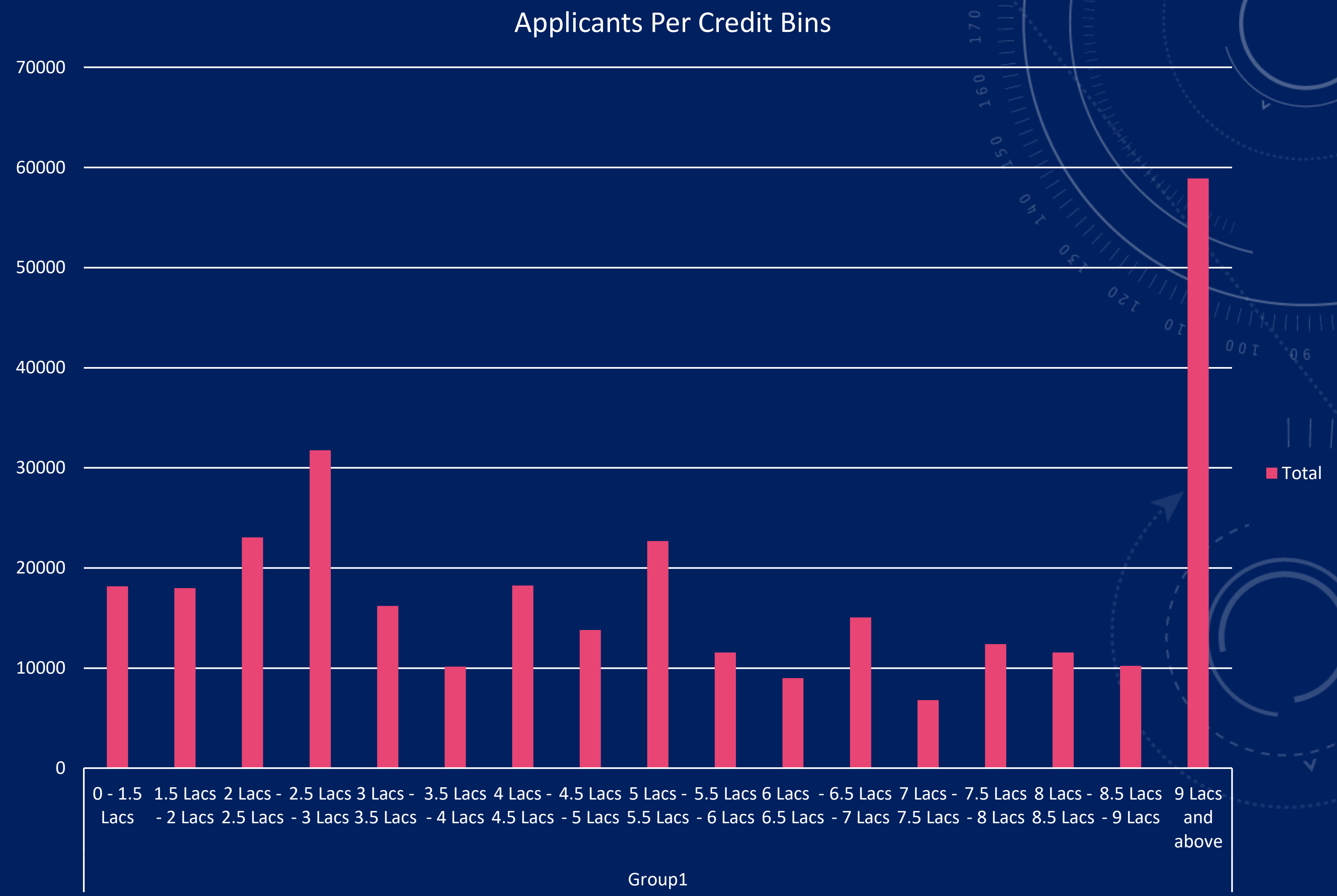
## Insights

In this "Data Imbalance" pie-chart indicates a ratio of 11.42 between applicants with payment difficulties (1) and those who paid installments on time (0). Out of 3,075,011 total applications, 92% of applicants paid their installments on time, making this the majority class. The remaining 8% of applicants had payment difficulties, constituting the minority class

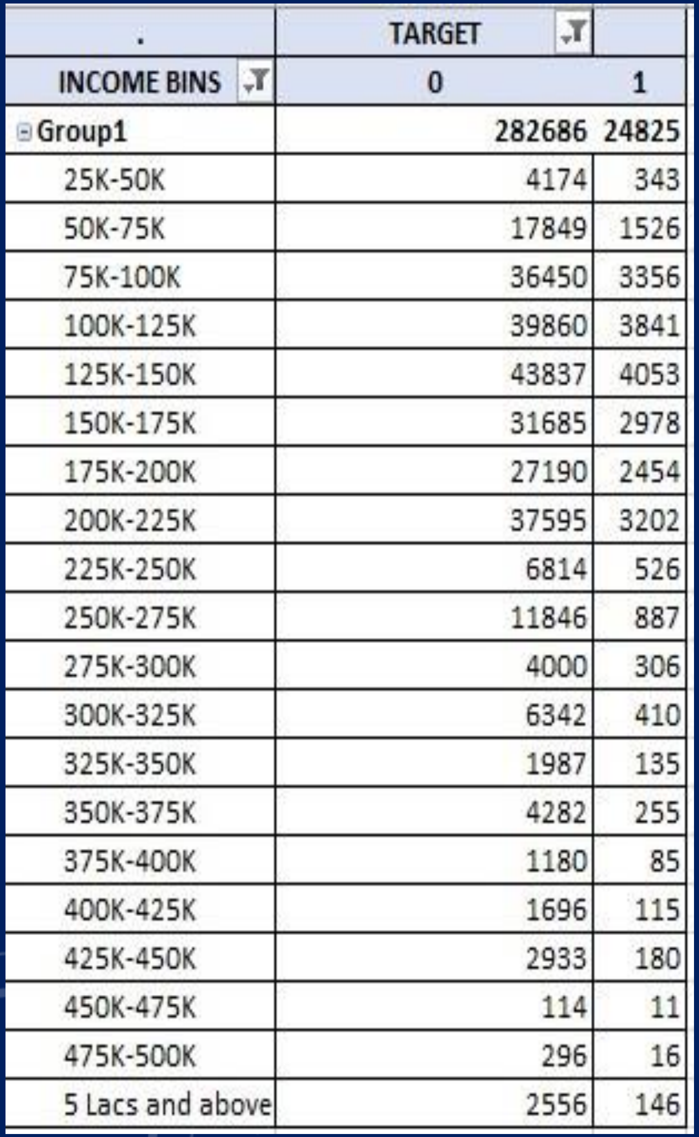
# Task 4.1:- Univariate Analysis



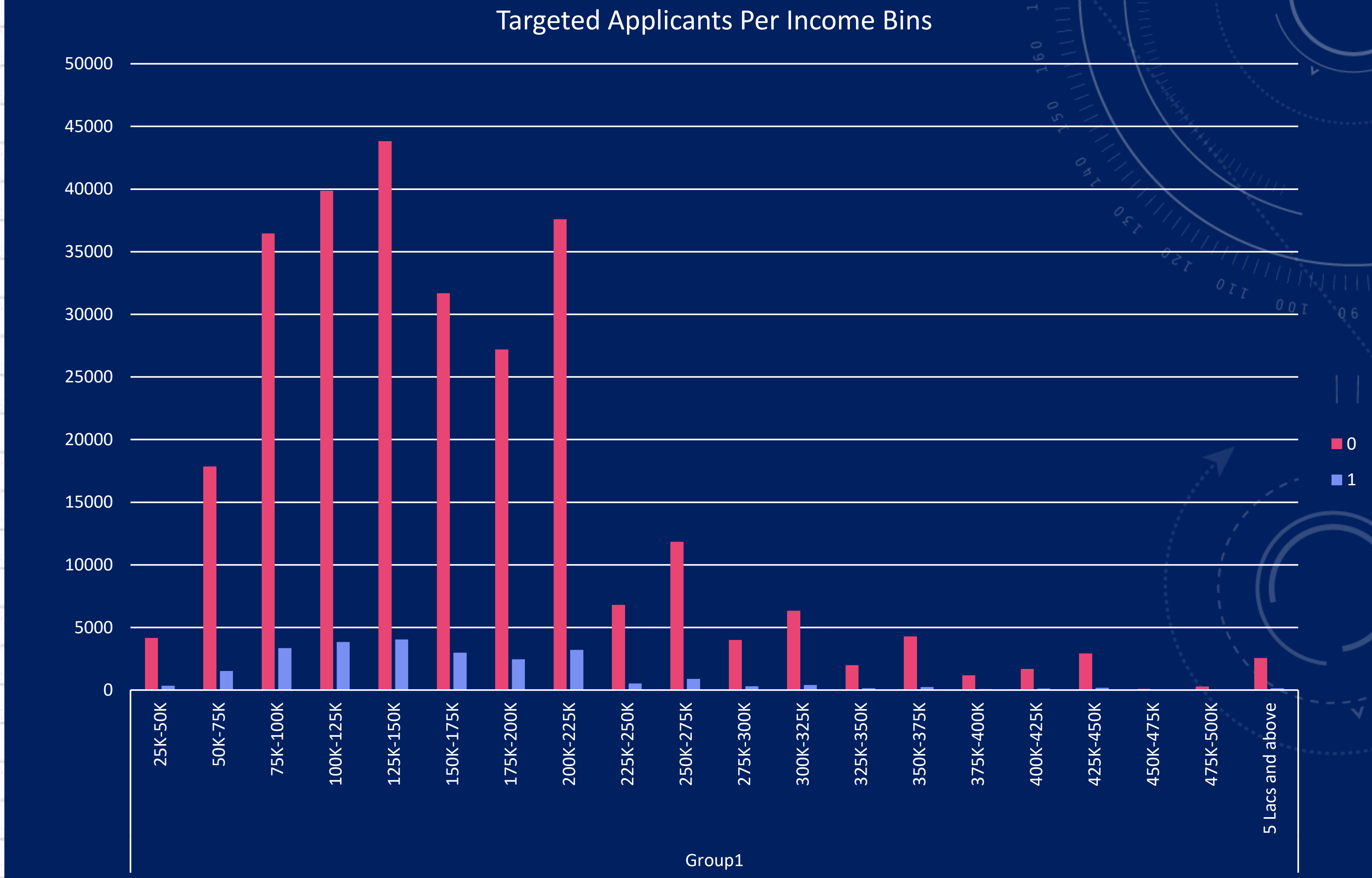
| CREDIT BINS | APPLICANTS |
|---|---|
| ⊟Group1 | 307511 |
| 0 - 1.5 Lacs | 18159 |
| 1.5 Lacs - 2 Lacs | 17985 |
| 2 Lacs - 2.5 Lacs | 23054 |
| 2.5 Lacs - 3 Lacs | 31759 |
| 3 Lacs - 3.5 Lacs | 16205 |
| 3.5 Lacs - 4 Lacs | 10133 |
| 4 Lacs - 4.5 Lacs | 18239 |
| 4.5 Lacs - 5 Lacs | 13799 |
| 5 Lacs - 5.5 Lacs | 22678 |
| 5.5 Lacs - 6 Lacs | 11554 |
| 6 Lacs - 6.5 Lacs | 8998 |
| 6.5 Lacs - 7 Lacs | 15051 |
| 7 Lacs - 7.5 Lacs | 6813 |
| 7.5 Lacs - 8 Lacs | 12380 |
| 8 Lacs - 8.5 Lacs | 11559 |
| 8.5 Lacs - 9 Lacs | 10233 |
| 9 Lacs and above | 58912 |



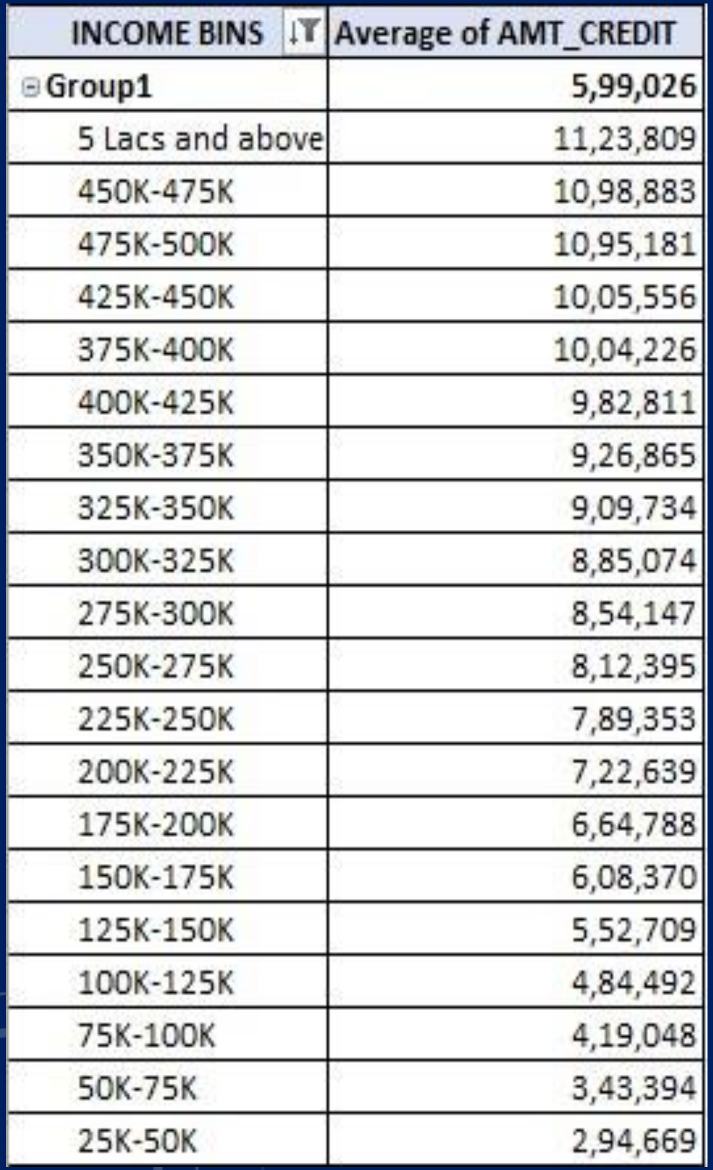
# Task 4.2:- Segmented Univariate Analysis



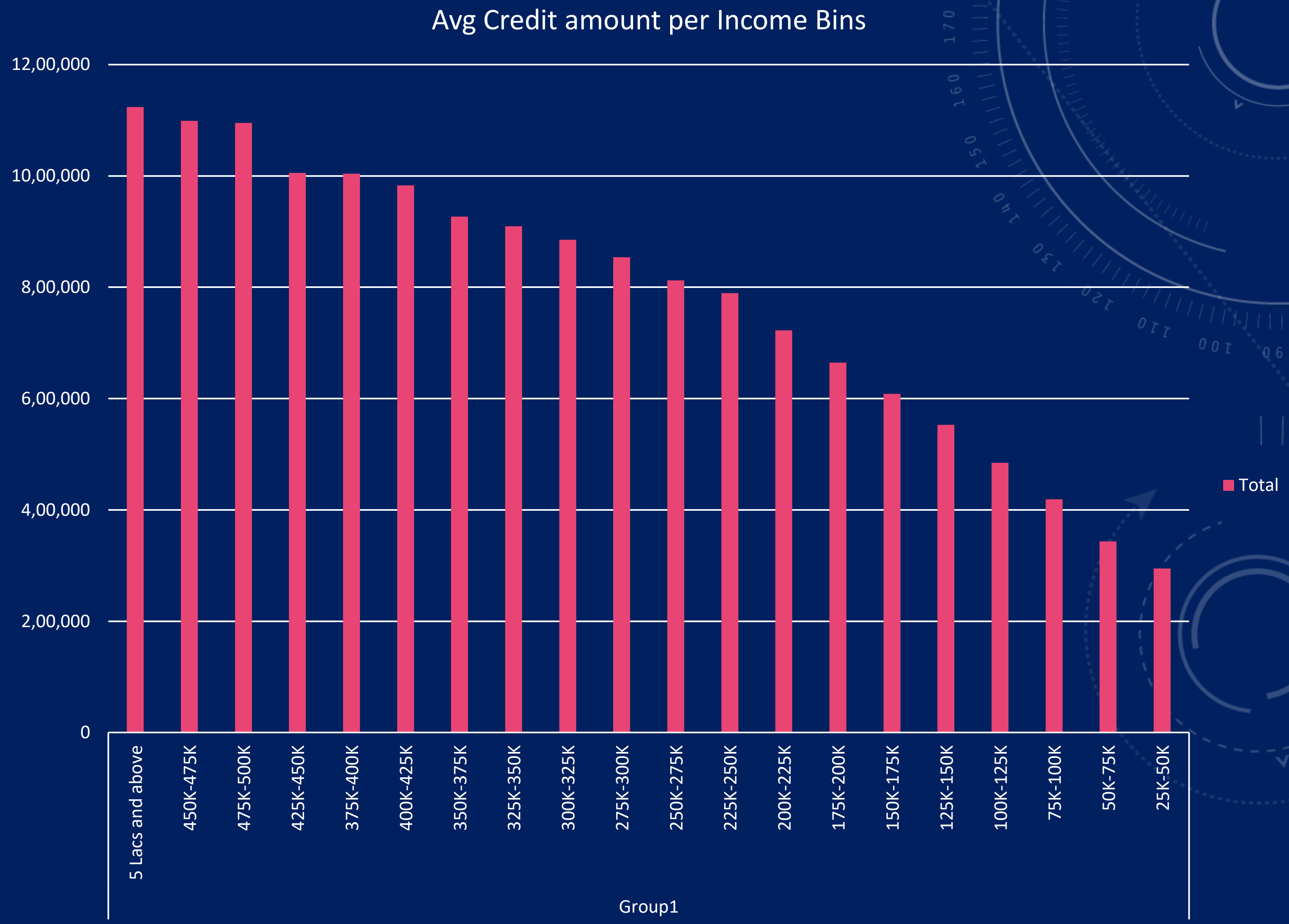
| . | TARGET | |
|---|---|---|
| INCOME BINS | 0 | 1 |
| ⊟Group1 | 282686 | 24825 |
| 25K-50K | 4174 | 343 |
| 50K-75K | 17849 | 1526 |
| 75K-100K | 36450 | 3356 |
| 100K-125K | 39860 | 3841 |
| 125K-150K | 43837 | 4053 |
| 150K-175K | 31685 | 2978 |
| 175K-200K | 27190 | 2454 |
| 200K-225K | 37595 | 3202 |
| 225K-250K | 6814 | 526 |
| 250K-275K | 11846 | 887 |
| 275K-300K | 4000 | 306 |
| 300K-325K | 6342 | 410 |
| 325K-350K | 1987 | 135 |
| 350K-375K | 4282 | 255 |
| 375K-400K | 1180 | 85 |
| 400K-425K | 1696 | 115 |
| 425K-450K | 2933 | 180 |
| 450K-475K | 114 | 11 |
| 475K-500K | 296 | 16 |
| 5 Lacs and above | 2556 | 146 |



# Task 4.3:- Bivariate Analysis



| INCOME BINS | Average of AMT_CREDIT |
|---|---|
| ⊟ Group1 | 5,99,026 |
| 5 Lacs and above | 11,23,809 |
| 450K-475K | 10,98,883 |
| 475K-500K | 10,95,181 |
| 425K-450K | 10,05,556 |
| 375K-400K | 10,04,226 |
| 400K-425K | 9,82,811 |
| 350K-375K | 9,26,865 |
| 325K-350K | 9,09,734 |
| 300K-325K | 8,85,074 |
| 275K-300K | 8,54,147 |
| 250K-275K | 8,12,395 |
| 225K-250K | 7,89,353 |
| 200K-225K | 7,22,639 |
| 175K-200K | 6,64,788 |
| 150K-175K | 6,08,370 |
| 125K-150K | 5,52,709 |
| 100K-125K | 4,84,492 |
| 75K-100K | 4,19,048 |
| 50K-75K | 3,43,394 |
| 25K-50K | 2,94,669 |



# Insights

1. **Univariate Analysis:-**

Univariate analysis involves examining data that contains a single variable. It does not focus on causes or relationships but rather aims to describe the data and identify patterns within it. The 1st graph above exemplifies univariate analysis, showing the count of applicants for the variable AMT_CREDIT grouped into different credit bins. Most applicants were offered loans in the credit range of 9 lakhs and above.

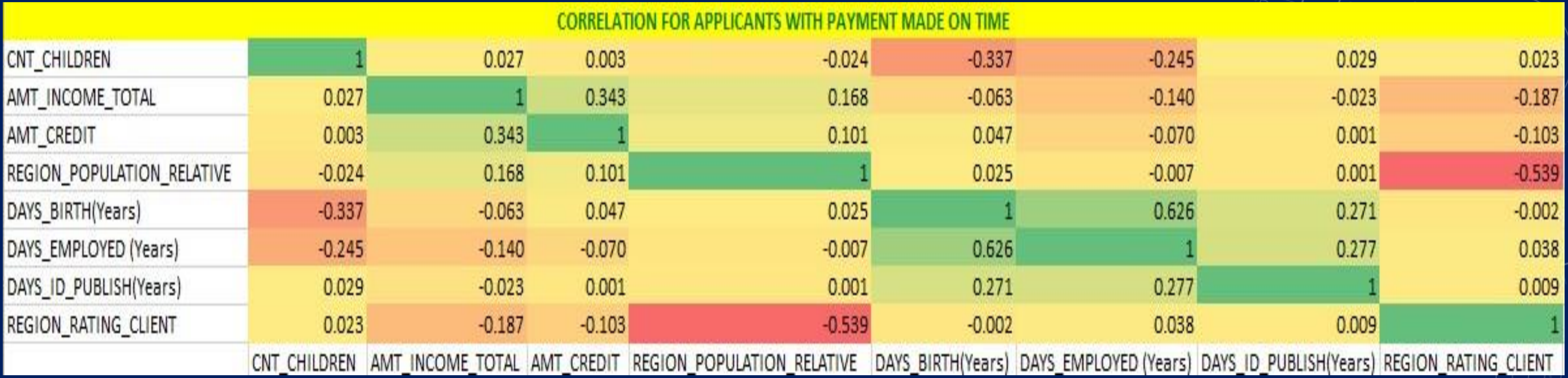
2. **Segmented Univariate Analysis:-**

Univariate analysis refers to examining data that contains only one variable. Segmented analysis means analyzing the data variable in subsets. The 2nd graph above illustrates univariate segmented analysis, showing the count of segmented applicants (0 and 1) for the variable AMT_TOTAL_INCOME, grouped into different income bins. The graph reveals that very few target 1 applicants earn more than 50 lakhs, which may contribute to their payment difficulties. Additionally, the majority of applicants (both 0 and 1) earn between 1.25 lakhs and 1.5 lakhs, though some within this income range still experience payment difficulties.

3. **Bivariate Analysis:-**

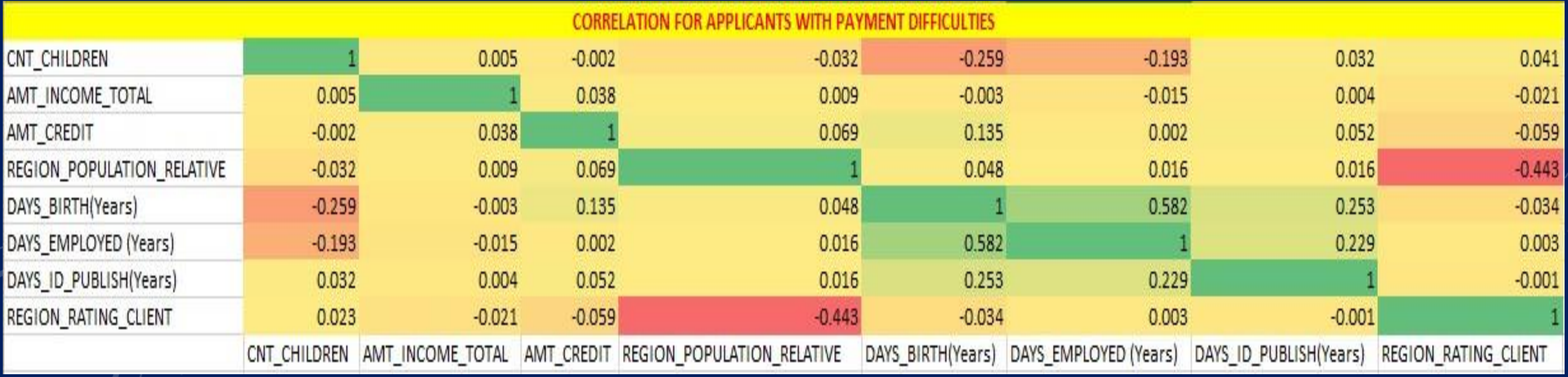
Bivariate analysis involves examining data that contains two variables, focusing on causes and relationships to identify how the variables interact. The 3rd graph above exemplifies bivariate analysis, showing the relationship between AMT_CREDIT and AMT_TOTAL_INCOME. The graph clearly indicates that applicants with higher incomes were offered higher loan amounts, demonstrating a direct proportional relationship between these two variables.

# Task 5:- Correlations for Different Scenarios



CORRELATION FOR APPLICANTS WITH PAYMENT MADE ON TIME

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| CNT_CHILDREN | 1 | 0.027 | 0.003 | -0.024 | -0.337 | -0.245 | 0.029 | 0.023 |
| AMT_INCOME_TOTAL | 0.027 | 1 | 0.343 | 0.168 | -0.063 | -0.140 | -0.023 | -0.187 |
| AMT_CREDIT | 0.003 | 0.343 | 1 | 0.101 | 0.047 | -0.070 | 0.001 | -0.103 |
| REGION_POPULATION_RELATIVE | -0.024 | 0.168 | 0.101 | 1 | 0.025 | -0.007 | 0.001 | -0.539 |
| DAYS_BIRTH(Years) | -0.337 | -0.063 | 0.047 | 0.025 | 1 | 0.626 | 0.271 | -0.002 |
| DAYS_EMPLOYED (Years) | -0.245 | -0.140 | -0.070 | -0.007 | 0.626 | 1 | 0.277 | 0.038 |
| DAYS_ID_PUBLISH(Years) | 0.029 | -0.023 | 0.001 | 0.001 | 0.271 | 0.277 | 1 | 0.009 |
| REGION_RATING_CLIENT | 0.023 | -0.187 | -0.103 | -0.539 | -0.002 | 0.038 | 0.009 | 1 |
| | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | REGION_POPULATION_RELATIVE | DAYS_BIRTH(Years) | DAYS_EMPLOYED (Years) | DAYS_ID_PUBLISH(Years) | REGION_RATING_CLIENT |



CORRELATION FOR APPLICANTS WITH PAYMENT DIFFICULTIES

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| CNT_CHILDREN | 1 | 0.005 | -0.002 | -0.032 | -0.259 | -0.193 | 0.032 | 0.041 |
| AMT_INCOME_TOTAL | 0.005 | 1 | 0.038 | 0.009 | -0.003 | -0.015 | 0.004 | -0.021 |
| AMT_CREDIT | -0.002 | 0.038 | 1 | 0.069 | 0.135 | 0.002 | 0.052 | -0.059 |
| REGION_POPULATION_RELATIVE | -0.032 | 0.009 | 0.069 | 1 | 0.048 | 0.016 | 0.016 | -0.443 |
| DAYS_BIRTH(Years) | -0.259 | -0.003 | 0.135 | 0.048 | 1 | 0.582 | 0.253 | -0.034 |
| DAYS_EMPLOYED (Years) | -0.193 | -0.015 | 0.002 | 0.016 | 0.582 | 1 | 0.229 | 0.003 |
| DAYS_ID_PUBLISH(Years) | 0.032 | 0.004 | 0.052 | 0.016 | 0.253 | 0.229 | 1 | -0.001 |
| REGION_RATING_CLIENT | 0.023 | -0.021 | -0.059 | -0.443 | -0.034 | 0.003 | -0.001 | 1 |
| | CNT_CHILDREN | AMT_INCOME_TOTAL | AMT_CREDIT | REGION_POPULATION_RELATIVE | DAYS_BIRTH(Years) | DAYS_EMPLOYED (Years) | DAYS_ID_PUBLISH(Years) | REGION_RATING_CLIENT |

# Insights

## 1. Correlations for Applicants with Payments Made on Time:

The 1st visualization in the above slide illustrates the correlations between various variables for the target group (0), representing applicants with no payment difficulties.
The most relevant correlations between the variables are:

- AMT_TOTAL_INCOME to AMT_CREDIT
- DAYS_EMPLOYED to DAYS_BIRTH
- REGION_POPULATION_RELATIVE to AMT_INCOME_TOTAL

## 2. Correlations for Applicants with Payment Difficulties:

The 2nd visualization displayed above illustrates the relationships between various variables for the target group (1), representing applicants experiencing payment difficulties.
Key correlations observed include:

- AMT_TOTAL_INCOME to AMT_CREDIT
- DAYS_EMPLOYED to DAYS_BIRTH
- REGION_POPULATION_RELATIVE to AMT_INCOME_TOTAL

# CONCLUSION

This project demonstrates effective techniques for handling large datasets, particularly through exploratory data analysis (EDA). When dealing with large datasets, it is crucial to selectively choose columns that are most pertinent to our analysis. Identifying correlated columns can significantly streamline this process, saving time and resources. Additionally, this project enhances understanding of key banking terminology.

Insights from the project include:

•Applicants with higher incomes were generally offered larger loan amounts by the bank.
•The majority of applicants and defaults had incomes ranging between 1.25 Lakhs and 1.5 Lakhs.
•A significant number of applicants were offered loans in the credit range of 9 Lakhs and above.

**Dataset:-** Bank Loan Case Study Analysis

**Loom Video Presentation:-** Bank Loan Case Study Analysis Presentation

# Thank You